ChinaXiv合作期刊
知识管理论坛 ISSN 2095-5472 CN11-6036/C
Knowledge Management Forum
E-mail: kmf@mail.las.ac.cn
http://www.kmf.ac.cn

【学术探索】

计算机科学领域科研合著网演化分析

◎ 邹鼎杰

南京政治学院上海校区军事信息管理系 上海 200433

摘要：[**目的 / 意义**] 揭示计算机科学领域科研合作发展规律及特点。[**方法 / 过程**] 以计算机科学领域 1975—2014 年约 300 万篇论文为基础构建合著网，以年为单位分析合著网演化特点，对比分析期刊论文合著网和会议论文合著网的特点。[**结果 / 结论**] 科研合作已经成为计算机科学领域的必然趋势和普遍现象，八成以上科研人员拥有两名或两名以上合作者；计算机科学领域的发展以 2000 年为界分为慢速发展和快速发展两个阶段；会议作为更为高效的科研信息交流方式，更有利于促进计算机科研领域的科学合作。

关键词：计算机科学 合著网 演化分析

分类号：G250

引用格式：邹鼎杰. 计算机科学领域科研合著网演化分析 [J/OL]. 知识管理论坛 , 2016, 1(2): 130-135[引用日期]. http://www.kmf.ac.cn/paperView?id=23.

随着科学研究的进一步发展，科研合作已经成为科学领域的普遍现象。科研论文是科研成果的主要表现形式，科研论文的合著能够从一定程度上反映科研合作状况。通过研究论文合著情况来了解科研合作现状，发现科研合作规律及特点，能够帮助科研管理人员加强科研管理，启发科研人员更好地开展科研合作。

国外的科学计量学专家普赖斯 [1] 和 D. Beaver 等 [2] 最早对科学合作和科研论文的联名发表情况进行了探究，普赖斯认为大多数高产作者提高他们的生产率是由于存在一个集体的领导而使他们能比单枪匹马时完成更多的工作量所致。国内的文献计量学专家王崇德 [3]、汪冰 [4] 等运用统计分析方法对合著率、合著程度等指标展开研究，通过量化的方法研究科研论文合著现象。随后 M.E. Newman[5]、A.L. Barabási 等 [6] 提出用网络方法研究合著关系，基于社会网络分析方法的合著现象研究越来越受到重视。王福生、杨洪勇 [7] 基于 2001-2006 年期间发表于《情报学报》的科学研究论文作者数据构建合著网络，探索了该网络的小世界特性和无标度特性；李亮、朱庆华 [8-9] 以 1998-2005 年期间发表在《情报学报》上的 880 篇科研论文为基础构建合著网络，利用社会网络分析方法对该合著网展开中心性分析、凝聚子群分析和边缘 - 核心结构分析；随后刘蓓、袁毅等 [10] 以 2000-2006 年被收录在中国期刊网上“情报学”主题相关的 9 806 篇论文为样本建立合作网络，分析研究人员合作紧密程度等特性。

然而，现有研究对象大多集中在图情学领域且数据规模小、时间跨度窄，通常以静态方式分析，缺乏动态分析研究。本文以计算机科

作者简介：邹鼎杰 (ORCID：0000-0001-6427-4519)，讲师，硕士，E-mail: yjqzou@163.com。

收稿日期：2015-11-30 发表日期：2016-04-19 本文责任编辑：王传清

知识管理论坛，2016(2):130-135
DOI：10.13266/j.issn.2095-5472.2016.017

学领域 1975-2014 年发表的约 300 万篇论文为样本建立合著网络，以年份为单位分析该网络演化特点，对比分析期刊合著网络和会议合著网络差异，揭示计算机科学领域的发展规律及特点。

1 DBLP 文献库及预处理

DBLP[11] 是由德国特里尔大学开发和维护的计算机科学文献库，该文献库收录了计算机科学领域主要的国际期刊和会议论文，为计算机科研人员提供免费的文献检索服务。由于其更新速度快，收录论文质量高，很好地反映了计算机领域学术前沿方向，在学术界有很好的声誉，给计算机科研人员带来了极大的便利，其权威性也得到了研究界的高度认可 [12]。截至 2015 年 8 月，该文献库已经收录超过 140 万名科研人员发表的约 360 万篇文献，其中期刊文献约 120 万篇，占 46%；会议论文约 160 万篇，占 53%。本文提取数据集中 1975-2014 年 40 年间发表的期刊论文和会议论文作为研究对象。

DBLP 数据集以 XML 格式提供数据服务，每条数据记录包含论文标题、作者、发表刊物、发表日期等字段。期刊论文以 <article> 节点标记，包含创建时间 (mdate) 和唯一标识 (key) 两个属性，以及作者（author）、标题（title）、刊名（journal）和发表年份（year）等子节点。一条典型的期刊论文记录属性见图 1。会议论文以 <inproceedings> 节点标记，所包含属性和子节点与期刊类似。由于各种原因，DBLP 文献库收录时存在部分期刊或会议论文字段不齐的情况。笔者挑选出创建时间、唯一标识、作者、标题、刊名和发表时间这 7 个要素均齐全的所有记录，删除 7 个要素不齐全的记录。最终得到 1975-2014 年间发表的 1 231 308 篇期刊论文和 1 607 382 篇会议论文。本文运用 java 语言，采用 sax 大规模 XML 文档处理程序包处理所有文档。

```
<article mdate=”2002-01-03” key=”persons/CoddD74”>
<author>E. F. Codd</author>
<author>C. J. Date</author>
<title>Interactive Support for Non-Programmers: The Relational and Network
Approaches.</title>
<journal>IBM Research Report, San Jose, California</journal>
<volume>RJ1400</volume>
<month>June</month>
<year>1974</year>
</article>
```

图 1 期刊论文典型记录

2 合著网络构建

本文主要考察科研作者之间有无合作关系，不考察合作关系强弱，因此建立无向无权值合著网络。以姓名为作者标识，作为合著网的节点；对于任意两名作者，如果他们合著过论文，则这两名作者之间建立一条无向边。最初以 1975 年发表论文为基础构建合著网，然后以 1975-1976 年间发表论文为基础构建合著网，以此类推，最终构建 1975-2014 年间发表的论文合著网络，分析这 40 年时间内随时间推演网络规模、度分布等演化特点。针对特定论文数据集，构架步骤如下：①基于论文数据构建“作者—合著者”关联表；②根据关联表统计当前合著网络规模；③根据关联表统计节点度及该网络度分布；④基于广度优先搜索算法分析该网络连通区域，并统计最大连通区域节点占整个网络的比例。

3 合著网演化分析

3.1 整体网络属性

表 1 显示了以 1975-2014 年间完整数据为基础构建的期刊合著网和会议合著网的基本属

性。期刊合著网作者人数约 93 万，共发表论文 123 万篇，平均每人发表论文 3.55 篇；会议合著网作者人数约 107 万，共发表论文 160 万篇，平均每人发表论文 4.52 篇。可以看出，计算机科研人员更倾向于以会议的形式发表科研成果，进行科研合作与交流。其原因是会议能够为计算机科研人员提供面对面交流机会，更有利于科研信息的快速交换，启发科学研究。会议合著网平均合作者为 7.73 人，高于期刊合著网的 6.90 人，说明科研人员在发表会议论文过程中更倾向于选择合著，这与会议论文的时效性和新颖性要求更高、同等质量论文需要更多科研人员参与才能完成有关。从连通性角度考虑，两者最大组元（组元指网络中的连通区域）节点数与网络总节点数比例均在 80% 以上，且第二大组元所占比例极低，说明合著网中除极个别的孤立节点外，绝大部分作者已经处于同一个连通区域，作者之间的联系越来越紧密；同时发现会议合著网的最大连通区域较期刊合著网大，会议论文的合著情况好于期刊论文。

表 1 期刊和会议合著网统计属性

属性	期刊合著网	会议合著网
论文数（篇）	1 231 308	1 607 382
作者数（人）	934 345	1 079 437
每名作者平均论文数（篇）	3.55	4.52
每名作者平均合作者数（人）	6.90	7.73
最大组元所占比例	81.847%	86.976%
第二大组元所占比例	3.8×10^{-5}	3.7×10^{-5}

3.2 网络规模演化分析

网络规模代表一个时期参与计算机科研的科研人员数量，科研人员数量的多少代表该领域受关注的程度。图 2 显示两个合著网历年新增节点数，横坐标为年份，纵坐标为当年新进入合著网的人数。从图 2 中可以看出，历年新加入计算机科学领域的研究人员呈上涨趋势，在信息化浪潮下，越来越多的科研人员加入到该领域研究中。根据新增的速度可以将计算机科学领域的发展分为两个阶段：第一个阶段为 1975-2000 年，网络规模缓慢增长，期刊网与会议网增长速度持平，呈现基本相当的趋势；第二个阶段为 2000-2014 年，在这 10 余年间网络规模数量迅猛增长，尤其是会议论文的增长量要高于期刊论文增长量。这主要得益于 2000 年前后，随着互联网等信息技术的兴起，计算机科学越来越受到重视，吸引了一大批科研人员参与该领域的研究，新增科研人员呈现爆发式增长。同时，由于计算机会议更有利于计算机科研人员的面对面交流且及时性更强，参与会议论文发表的科研人员人数一直多于期刊论文科研人员。

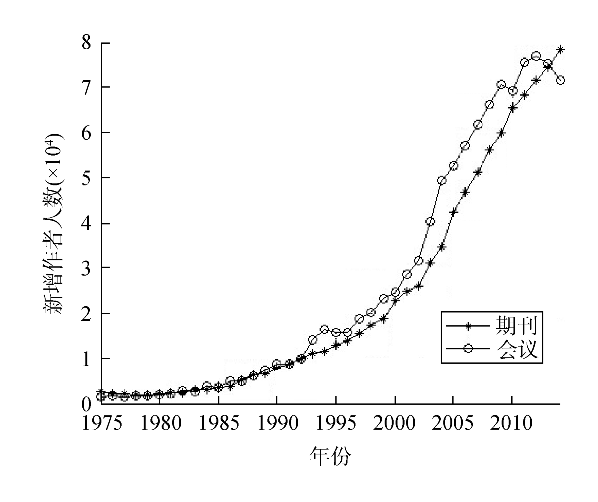


图 2 合著网规模演化分析

3.3 度分布演化分析

网络中节点的度指该节点的相邻节点数。节点度分布，是指度为 n 的节点数与整个网络节点数的比值。在合著网络中，一个节点代表一名科研人员，节点度代表该科研人员具有的合著者人数，网络的平均度代表该网络中平均每名科研人员拥有的合著人数。合著网的度分布代表拥有不同合作人数的科研人员分布，能够比平均度更为详细地反映该领域科研合著程

知识管理论坛，2016(2):130-135
DOI：10.13266/j.issn.2095-5472.2016.017

度。图 3 为两个网络平均度演化趋势，从图 3 中可以看出随着时间的演化，科研人员平均合作人数逐年呈线性增加。说明在计算机科学领域，合作已经成为科学研究的总体趋势，单独工作或者仅具有较少的合作者都难以高效完成工作。同时，通过对比期刊合著网和会议合著网可以发现，会议合著网中科研人员拥有的合作人数要多于期刊合著网。这主要是由于会议对论文的生产周期要求更短，完成同等质量的科研论文需要更多的科研人员共同参与。

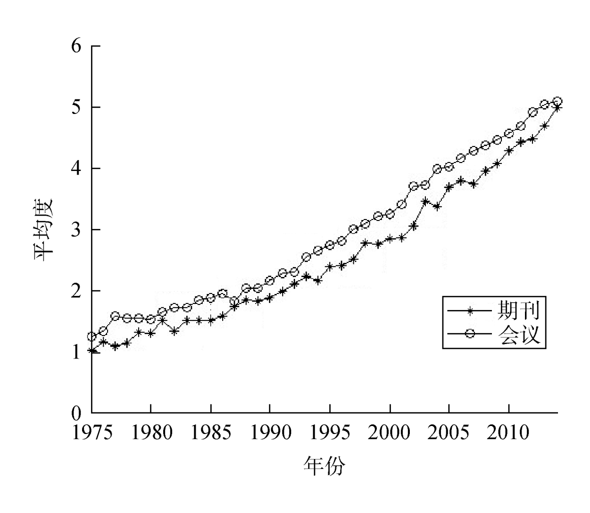


图 3 期刊和会议合著网平均度演化分析

平均度代表该领域科研人员拥有合作者的平均数量，度分布能够更详细地呈现科研人员拥有的合作人员情况。从整体情况来看，80% 以上科研人员拥有的合作者人数在 6 人以内，本文重点考察度为 0-6 的节点分布情况。图 4 和图 5 分别为期刊合著网和会议合著网的度分布演化分析图，图的横坐标为年份，纵坐标为该年合著网中度为 n 的节点数占整个网络节点的比例。从整体趋势可以看出，度为 0 和度为 1 的节点数逐年下降，度为 2 的节点比例经历一段上升之后也从 2000 年开始下降，度为 3-6 的节点比例逐年上升且势头明显。上述现象表明独立完成论文或者仅有一名合作者的科研人员越来越少，越来越多的科研人员拥有两名以及两名以上的合作者，合作研究已经成为当前计算机科学研究领域的主流现象。从定量分析可以看出，期刊合著网中独立完成论文的科研人员比例已经下降到 2014 年的 3%，会议合著网中这一比例更是下降到 2%，说明独立完成论文的情况已经非常稀有。拥有 2-4 名合作者的科研人员比例超过了 50% 且有继续上升的趋势，拥有 2-4 名合作者已经成为计算机科学研究领域的主流现象；拥有 5-6 名合作者的科研人员逐年上升，说明当前趋势下，大规模合作将成为未来计算机科学研究领域的趋势之一。

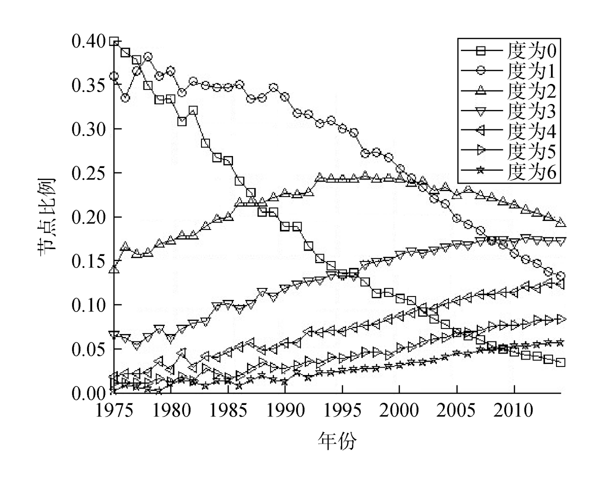


图 4 期刊合著网度分布演化分析

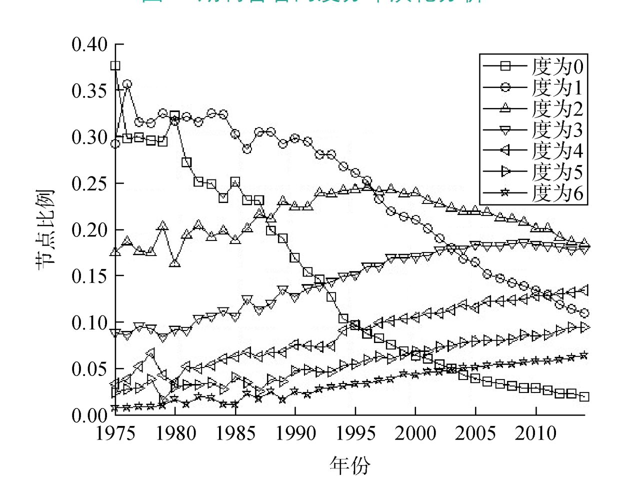


图 5 会议合著网度分布演化分析

3.4 最大组元演化分析

在社会网络形成初期，由于各个节点间

的连接较少，网络通常呈现出比较分散的孤立节点或者规模较小的孤立区域；随着时间的推移，网络中原有节点之间加强交流进一步建立连接，新加入节点也将与原有节点建立连接，网络将逐渐演化成一个巨大的连通区域，以及若干个非常小的孤立区域。通常将这些区域称作组元，一个组元代表网络中的一个社区。下面将考察在计算机科学领域科研合著网中的组元演化现象。

计算机科学领域合著网络倾向于连接成一个整体网络。在网络发展的初期，网络由较多的小型组元构成，呈现出较为分散的状态；随着时间演化，组元之间逐渐连接成更大组元，网络最终由一个规模巨大的组元以及若干个规模非常小的组元构成。期刊合著网在 1975 年包含 1 718 个组元，最大组元节点数占整个网络节点比例为近 0.48%，第二大组元比例为 0.47%，约 200 个组元的规模都在 0.1% 以上，网络特点表现为由较多小规模组元构成，组元之间呈现孤立的分散状态。期刊合著网发展至 2014 年，最大组元比例上升至 81%，第二大组元比例为 2.5×10^{-5}，已经呈现出绝大多数科研人员形成一个的巨型组元、零星的独立科研人员形成微小组元的特点。这说明随着该学科的发展和演化，科研人员之间的合作关系倾向于越来越密切，科研合作朝着正常的方向发展。

图 6 显示了期刊合著网和会议合著网最大组元相对大小随时间变化的趋势。从图 6 可以看出最大组元由最初的 3% 左右，逐渐演化到规模在 80% 以上。在网络发展后期，最大组元通常会在 90% 处缓慢增长，总会存在约 10% 的孤立节点不与最大组元形成连接。笔者认为，缓慢增长主要源于以下两个原因：第一个原因是存在仅发表一篇论文的独立作者，这类作者每年都会新产生一部分，为整个网络的永久性孤岛，这就导致了理论上合作网络永远不可能形成一个完全连通的网络；另一个原因是，部分新加入节点无法在当年就与最大组元建立连接，而形成一个动态的孤立区域。

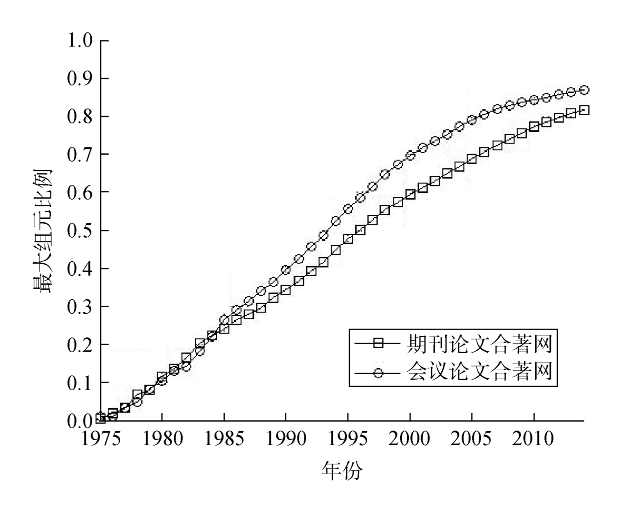


图 6 最大组元比例演化分析

4 结论及启示

本文以 DBLP 数据集中 1975-2014 年间 300 万篇论文数据为基础构建科研合作网络，以年份为序分析计算机科学领域科研合著网络的发展特点，从合著网络基本属性、网络规模、度分布以及最大组元等 4 个角度揭示计算机领域合著现象发展规律及特征。

计算机科学领域合著现象表现为以下 3 个特征：①合著已成为该领域的普遍现象和必然趋势，科研合作是该领域科学研究的主要方式。从度分布演化分析可以看出，该领域科研人员的合作人数逐年上升，至 2014 年，独立完成论文的科研人员比例已经下降至 3%，拥有 2-4 名合作者的科研人员比例超过 50% 且有继续上升的趋势。最大组元演化形成一个巨型的连通网络也反映出该领域科研人员的合作关系越来越密切。②计算机科学领域的发展分为两个阶段。第一个阶段为 1975-2000 年，网络规模缓慢增长；第二个阶段为 2000-2014 年，网络规模数量迅猛增长。③会议论文相对于期刊论文更能够促进科研人员参与合作。从期刊论文合著网和会议论文合著网的网络规模增长曲线可以看出，会议合著网的增速高于期刊合著网，会议吸引了更多的科研人员参与；从最大组元演化分析可以看出，会议合著网最大组元增速在 2000 年以后一直高于期刊论文合著网的最大组元，说明会议论

文形成的合作关系更为密切，会议论文平均度一直高于期刊论文也说明了这一点。

本文认为产生上述特征的因素可能有以下3个：第一，科研难度增加以及科学家乐于合作共同解决难题是促进科研合作的主要原因。在一个研究领域发展早期，科研人员倾向于解决基本问题，这类问题比较直观、所需投入的人力较少；随着基本问题的解决，复杂难题需要科学家付出更多的时间和精力，而人类乐于合作的天性也促使科学家走在一起，以更为高效的方式共同解决难题。第二，2000年以来计算机和互联网市场的迫切需求刺激了该领域的发展。通过两类合作网网络规模的增长可以明显看出，2000年以后网络规模呈现出爆发式增长趋势，越来越多的计算机科研人员参与到该领域的研究中。本文认为出现这种明显变化的原因是2000年左右互联网和计算机技术的蓬勃发展吸引了一大批人员参与到计算机科学领域的研究中。第三，会议对论文的时效性要求更高，同等质量的科研论文需要在更短时间内完成，这样从客观上要求科研人员加强合作，提高科研效率；同时会议能够为科研人员提供面对面形式的科研信息交流，可能是吸引更多科研人员参与会议的原因之一。

基于海量数据的合著网分析能更加准确、全面地呈现一个学科合作的发展现状，但因为面临着数据处理难题，传统的个人电脑几乎无法完成一些常见指标（如网络直径等典型参数）的计算。在下一步工作中，笔者将进一步探索如何高效地进行海量数据处理和巨型合著网络的分析和计算。

参考文献：

[1] 普赖斯 . 小科学 , 大科学 [M]. 宋剑耕 , 戴振飞 , 译 . 北京 : 世界科学社 , 1982.

[2] BEAVER D, ROSEN R. Studies in scientific collaboration: part I. the professional origins of scientific co-authorship[J]. Scientometrics, 1978, 1(1): 65-84.

[3] 王崇德 . 科学论文作者的研究 [J]. 情报学报 , 1982, 1(2): 220-225.

[4] 汪冰 . 我国自然科学期刊论文合著现象研究 [J]. 情报学刊 , 1990, 11(5): 335-339.

[5] NEWMAN M E. The structure of scientific collaboration networks[J]. Working papers, 2000, 98(2): 404-409.

[6] BARABÁSI A L, JEONG H, NÉDA Z, et al. Evolution of the social network of scientific collaborations[J]. Physica A: statistical mechanics and its applications, 2002, 311(3-4): 590-614.

[7] 王福生 , 杨洪勇 .《情报学报》作者科研合作网络及其分析 [J]. 情报学报 , 2007, 26(5): 659-663.

[8] 李亮 , 朱庆华 . 社会网络分析方法在合著分析中的实证研究 [J]. 情报科学 , 2008(4): 549-555.

[9] 朱庆华 , 李亮 . 社会网络分析法及其在情报学中的应用 [J]. 情报理论与实践 , 2008, 31(2): 179-183.

[10] 刘蓓 , 袁毅 , BOUTIN E. 社会网络分析法在论文合作网中的应用研究 [J]. 情报学报 , 2008, 27(3): 407-417.

[11] [EB/OL]. [2015-11-08]. http://dblp.uni-trier.de/db/.

[12] 窦炳琳，李澍淞，张世永 . 基于结构的社会网络分析 [J]. 计算机学报 , 2012, 35(4): 741-753.

Analysis on the Evolution of Author Cooperative Network in Computer Science

Zou Dingjie
Department of Military Information Management, Shanghai Branch of Nanjing Institute of Politics, Shanghai 200433

Abstract: [Purpose/significance] This paper aims at finding out the characteristics of author cooperation in computer science. **[Method/process]** We built the author cooperative network with about 300 minions of papers from 1975 to 2014, and we analyzed the evolution of this network. We built two networks with papers in journals and papers in conferences, and compared the differences of two types of networks. **[Result/conclusion]** Co-author is a universal phenomenon, and about 80% scientists have more than three cooperators. There are two stages in computer science, the slowly developing stage before 2000 and the rapid stage after 2000. It is found that scientists are more cooperative in conferences than in journals.

Keywords: computer science cooperative network evolution analysis